The Quarterly Journal of Experimental Psychology

ISSN: 1747-0218 (Print) 1747-0226 (Online) Journal homepage: http://www.tandfonline.com/loi/pqje20

Causal judgements about temporal sequences of events in single individuals

Peter A. White

To cite this article: Peter A. White (2015) Causal judgements about temporal sequences of events in single individuals, The Quarterly Journal of Experimental Psychology, 68:11, 2149-2174, DOI: 10.1080/17470218.2015.1009475

To link to this article: http://dx.doi.org/10.1080/17470218.2015.1009475

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
Vol. 68, No. 11, 2149–2174, http://dx.doi.org/10.1080/17470218.2015.1009475

Causal judgements about temporal sequences of events in single individuals

Peter A. White

School of Psychology, Cardiff University, Cardiff, UK

(Received 27 August 2013; accepted 7 January 2015; first published online 2 March 2015)

Stimuli were presented in which values of an outcome variable for a single individual were recorded over 24 time periods, and an intervention was introduced at one of the time periods. Participants judged whether and how much the intervention affected the outcome. Judgements were affected by manipulations of the temporal relation between the intervention and a gradual increase in values on the outcome variable, by the size of the increase, by the time taken for the increase to occur, and by variance in the preincrease data. Most results were predicted by a simple model in which the mean outcome value for the preintervention time periods is subtracted from the mean outcome value for the postintervention time periods, though there was also an effect of temporal contiguity that is not predicted by the simple model. This form of information, which is a kind of quasiexperimental design, is more representative of the kind of information generally available for causal judgement than the more commonly investigated binary variables in which the cause is either present or absent, and the outcome either occurs or does not; as such, it is more revealing of how causal judgements are made under the conditions that prevail in the world.

Keywords: Causal judgement; Quasiexperimental design.

Causal judgements about temporal sequences of events in single individuals

Some years ago I suffered a back injury and saw a physiotherapist. The physiotherapist was at first unwilling to treat me because we did not have a diagnosis. While waiting for a diagnosis, for several months, my condition did not improve. When the diagnosis arrived, the physiotherapist started me on some exercises. During that period, my condition improved slowly but steadily, again over a period of many months. As a result of that, I acquired a belief that the physiotherapy exercises played a causal role in my improvement. An article in New Scientist magazine reviewed the use of transcranial magnetic stimulation (TMS) to treat patients in a persistent vegetative state. Discussing one such patient, the article said:

> He had only been given a 20 to 40 per cent chance of long-term recovery, and until he was given TMS his functioning had not improved since about four months after the accident. What's more, after the 15th TMS session, he improved incrementally with each session—further evidence that TMS was the cause. (Geddes, 2008, p. 9)

What these anecdotes have in common is the observation of incremental change from an established baseline, temporally associated with an intervention, with a single participant. Causal inferences from such data cannot be scientifically justified, for two obvious reasons. First, the data come from a sample of one. Except in rare cases where there is essentially zero variance in a population, causal

Correspondence should be addressed to Peter A. White, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3YG, UK. E-mail: whitepa@cardiff.ac.uk

inferences require large samples of observations. Testing the efficacy of physiotherapy exercises, for example, would require a large sample of patients with similar injuries performing similar exercises. Second, the data constitute, at best, a quasiexperimental design with no control group. Many other factors temporally correlated with the intervention could be responsible for the change, and the data provide no means of ruling any of them out (Cook & Campbell, 1979).

Despite this, causal inferences from such data are peculiarly persuasive, as I can testify from my own experience. I suspect that inferences of this kind are commonly made by laypeople. Probably most of us, for example, see some of our adult characteristics, habits, fears, and predilections as outcomes of events or extended occurrences in our early lives. We may trace incremental changes in health-related factors to changes in our diet, or improvement in our mental state to the adoption of an exercise programme. All of these kinds of causal inferences suffer from the same methodological inadequacies: We have a sample of one, and we cannot run a control condition. We do not know what our sample of one would have been like in the absence of the intervention. Laypeople do not have the opportunities that scientists have to collect large amounts of data under controlled experimental manipulations. The activity of causal inference in everyday life is inevitably methodologically flawed. If we want to make causal inferences about ourselves at all, we have no choice but to accept whatever data we can get.

Laboratory investigations of human causal inference have paid little attention to issues of the methodological adequacy of data that can be obtained under the conditions of everyday life. There have been many studies, for example, of causal inference from information about empirical associations between binary variables (Allan, 1993; De Houwer & Beckers, 2002; Hattori & Oaksford, 2007; Perales & Shanks, 2007). These are variables with two values: The cause in question may be present or absent, and the outcome may occur or not. Stimulus presentations are designed with at least a moderate sample of observations, and with information about what happens when the cause of interest is absent as well as what happens when it is present. The cause-absent information is usually presented in such a way as to constitute a methodologically appropriate control condition. These methodological features are employed in part because data from them can be analysed with normative inferential procedures, yielding either objective contingencies (Jenkins & Ward, 1965; McKenzie, 1994; Ward & Jenkins, 1965) or, so it has been claimed, normative causal analyses (Cheng, 1997; Griffiths & Tenenbaum, 2005, 2009; Holyoak & Cheng, 2011; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008).

In everyday life it would be impractical to forgo causal inference on grounds of methodological or informational imperfections; it is surely more likely that causal judgement is suited to the imperfect conditions of life than to the idealized conditions of the scientific laboratory. How do people make causal inferences under such conditions? In the present research, I attempt to shed some light on this by looking at the case described earlier: single instances with information about change over a series of time periods, associated with specific interventions. This is just one among many possible forms of data that may be available in the world, but I suggest that it is one of the more important for human causal judgement, not least because it relates to the problem of making self-relevant causal inferences.

There are several factors that could influence causal judgements in the kind of situation investigated here. The most obvious is temporal association. Many studies have shown that temporal cues are involved in causal inferences by children and adults and sometimes override other cues such as objective contingencies (Bühner & May, 2002, 2003, 2004; Bühner & McGregor, 2006; Greville & Buehner, 2007; Lagnado & Sloman, 2006; Mendelson & Shultz, 1976; Rottman & Keil, 2012; Schlottmann, 1999; Shultz & Kestenbaum, 1985; Siegler, 1975; Siegler & Liebert, 1974; White, 1988, 2006). The use of temporal information tends to interact with mechanism beliefs. If people believe that a mechanism connecting a cause to an outcome takes a certain amount of time to operate, they tend not to identify a

temporally proximal event as the cause if they believe there is insufficient time for the mechanism to operate (Bühner & McGregor, 2006). In the absence of relevant mechanism beliefs, the research cited above shows that temporal contiguity (along with temporal order) is a strong cue to causal inference. However, none of the studies has considered causal inference from a time series design concerning a single entity with no control condition.

Another possibly relevant factor is change magnitude. Incremental change is not a binary variable. I can assess the amount of improvement in my back, for example, compared to a baseline (e.g., the state it was in just before treatment started), and I can assess how that amount changes as time progresses. Change magnitude has received much less attention in causal judgement research than contingency information has, but there are indications that causal judgement is influenced by outcome magnitude information from an early age. In the first year of life, infants have expectations about the relation between the size of an object and the magnitude of outcome it should produce, and they are surprised when those expectations are violated (Kotovsky & Baillargeon, 1998). This suggests that, from an early age, we operate with a simple rule relating the strength of a cause to the magnitude of the outcome: The stronger the cause, the greater the outcome. Evidence consistent with the use of this rule has been reported (diSessa, 1993). If this is the case, then the amount of change in the value of a variable may be taken as an indicator of causal strength.

With incremental changes, two dimensions of change magnitude can be distinguished. One is the amount of change that occurs, either in absolute terms or relative to a baseline. An improvement of 50% in my back might make a more persuasive case for the efficacy of physiotherapy than an improvement of 25% would. The other is the slope of the improvement function. This means, in effect, the amount of time it takes to reach a given value of outcome magnitude. If we suppose, for the sake of simplicity, that the function is linear, then an improvement of 50% in three months is likely to be more persuasive than an improvement of 50% in six months.

Experiment 1 was therefore designed to investigate effects of temporal association, amount of change, and slope of the improvement function. Participants are told that scientists are monitoring the level of cells of a particular kind in the human bloodstream, and they want to see whether this quantity is affected by a chemical injected into the blood. The patient's blood is sampled once an hour for 24 hours. So there is information about each of 24 consecutive time periods. Each time period presents a record of the quantity of cells counted, and the point at which the intervention (the injection) is made is also indicated. This design is called a simple interrupted time series (Cook & Campbell, 1979), where the intervention is the interruption.

EXPERIMENT 1

Experiment 1 incorporated two experimental designs, hereafter Experiments 1a and 1b, in a single set of stimulus materials with the same sample of participants. In all, there were 20 individual judgemental problems, hereafter called datasets. The design of Experiment 1a incorporated 16 of these. The design of Experiment 1b incorporated four of those in the Experiment 1a design and four additional datasets. Some data, therefore, enter the analyses for both designs. Order of presentation of datasets was randomized independently for each participant.

EXPERIMENT 1A

Method

Participants

The participants were 39 first-year undergraduate students of psychology with English as their first language. They received course credit for their participation. None had been taught any psychology of relevance to this topic.

Stimulus materials

The materials comprised an initial instruction sheet and 16 datasets. The initial instructions read as follows:

The human bloodstream contains many different kinds of cells. Scientists are trying to find out what influences levels of these cells in the blood. They are testing the hypothesis that certain chemicals may have an effect on these cells. To test this, they monitor the level of a given kind of cell in the blood over a period of time; some time during that period they inject a chemical into the bloodstream to see what effect it has, if any.

On the following pages you will see data for a series of experimental trials of this sort. Each page concerns a different kind of cell and a different kind of chemical, each identified with a two-letter code. The data concern a single patient, identified by a number. The patient's blood is sampled once every hour for 24 hours, and the level of the cells in the patient's blood in each of these samples is given, in millilitres per litre of blood. So you will see two columns of information. The left-hand column just numbers the hours, in chronological order. The right-hand column gives the level of cells recorded for each hour. At some point during this sampling period the chemical in question is injected into the patient's blood. The data tell you at what point this was done on each page.

At the bottom of the page you will see two questions. The first question asks you whether the chemical causes an increase in the level of the cells in the patient's blood or not. You just answer yes or no to this. If you answer yes, then you go on to the second question. This asks you to rate how strong a cause of increase the chemical is. You should answer this question by writing a number from 1 to 100 beside it. 1 means that the chemical is a very weak cause of increase in the cells, and 100 means that it is a very strong cause of increase in the cells. The stronger you think the chemical is as a cause of increase in the cells, the higher the number you should put, up to a maximum of 100.

Following this were 20 datasets, each presented on a separate page, and each with the same format. Sixteen of these constituted the materials for Experiment 1a, and the other four, together with four of the Experiment 1a datasets, constituted the materials for Experiment 1b. At the top of each page the chemical and the cell type were identified with two-letter codes, and the patient was identified with a number. Each page had different identifiers, to make it clear that each dataset was independent of the others. Under that were the columns of information as described above. At the appropriate point the sentence "chemical injected here", in block capitals, was inserted between adjacent time periods. At the bottom of the page were the two questions, as described above.

The datasets are all shown in Table 1. Each column is a dataset, and the headings identify the values of each of the independent variables. The location of the intervention is indicated with an asterisk. In all datasets, readings in the first six time periods alternated between 15 and 16. That is the baseline phase. At time period 7 an incremental change began. That is the increase phase. The duration of the increase phase varied depending on whether the rate of increase was shallow (11 time periods) or steep (6 time periods). Once the peak value was reached, the remaining readings oscillated between that value and one lower. That is the postincrease phase.

Design

Three variables were manipulated, all within subjects. The timing of the intervention was manipulated. The intervention could occur three time periods before the start of the increase ("early"), in the time period where the increase began ("on time"; the location lies between time periods 6 and 7), midway between the beginning and the end of the increase ("midway"; the exact point of this depends on the duration of the increase phase), and in the time period where the increase ended ("late"; the exact point of this also depends on the duration of the increase phase).

Amount of increase, operationalized as the difference between the levels in the last and first of the increase periods, was manipulated with two values, small (5) and large (10). Rate of increase, operationalized as the duration of the increase phase, was manipulated with two values, shallow (11 time periods) and steep (6 time periods). The design was therefore $4 \times 2 \times 2$, totalling 16 datasets.

Procedure

Participants took part in small groups, supervised by an experimenter. The questionnaire for this experiment was included among a set of materials for experiments on unrelated topics. Participants were told that they should ask questions if anything in the instructions was not clear. None had any questions about the materials for this experiment. Participants then proceeded through the tasks at their own pace. At the end of the session, participants were given course credit and debriefed

Table 1. *The 16 datasets used in Experiment 1a*

Period	*e/s/ sh*	*e/s/ st*	*e/l/ sh*	*e/l/ st*	*o/s/ sh*	*o/s/ st*	*o/l/ sh*	*o/l/ st*	*m/s/ sh*	*m/s/ st*	*m/l/ sh*	*m/l/ st*	*l/s/ sh*	*l/s/ st*	*l/l/ sh*	*l/l/ st*
1	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
2	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
3	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
4	15*	15*	15*	15*	15	15	15	15	15	15	15	15	15	15	15	15
5	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
6	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
7	17	17	17	17	17*	17*	17*	17*	17	17	17	17	17	17	17	17
8	17	18	18	19	17	18	18	19	17	18	18	19	17	18	18	19
9	18	19	19	21	18	19	19	21	18	19	19	21	18	19	19	21
10	18	20	20	23	18	20	20	23	18	20*	20	23*	18	20	20	23
11	19	21	21	25	19	21	21	25	19	21	21	25	19	21	21	25
12	19	22	22	27	19	22	22	27	19	22	22	27	19	22*	22	27*
13	20	21	23	26	20	21	23	26	20*	21	23*	26	20	21	23	26
14	20	22	24	27	20	22	24	27	20	22	24	27	20	22	24	27
15	21	21	25	26	21	21	25	26	21	21	25	26	21	21	25	26
16	21	22	26	27	21	22	26	27	21	22	26	27	21	22	26	27
17	22	21	27	26	22	21	27	26	22	21	27	26	22*	21	27*	26
18	22	22	27	27	22	22	27	27	22	22	27	27	22	22	27	27
19	21	21	26	26	21	21	26	26	21	21	26	26	21	21	26	26
20	22	22	27	27	22	22	27	27	22	22	27	27	22	22	27	27
21	21	21	26	26	21	21	26	26	21	21	26	26	21	21	26	26
22	22	22	27	27	22	22	27	27	22	22	27	27	22	22	27	27
23	21	21	26	26	21	21	26	26	21	21	26	26	21	21	26	26
24	22	22	27	27	22	22	27	27	22	22	27	27	22	22	27	27

Note: Columns are identified by values of independent variables in the order timing/magnitude/rate. For timing, "e" = early, "o" = on time, "m" = midway, and "l" = late. For amount of increase, "s" = small and "l" = large. For rate of increase, "sh" = shallow and "st" = steep. Locations of interventions are indicated by asterisks.

about the general aims of the research, but not about the specific hypotheses being tested.

Results

For purposes of analysis, "no" responses to the first question were treated as ratings of zero. Data were analysed with a 4 (timing of intervention: early vs. on time vs. midway vs. late) × 2 (amount of increase: small vs. large) × 2 (rate of increase: shallow vs. steep) within-subjects analysis of variance (ANOVA). The Tukey test was used for post hoc paired comparisons in this and subsequent experiments. Means for the 16 datasets are presented in Table 2. There were significant effects of all three variables.

There was a significant main effect of timing, $F(3, 114) = 15.06$, $MSE = 416.37$, $p < .001$, $\eta_p^2 = .28$. Post hoc paired comparisons revealed a significantly lower mean for late intervention (10.44) than for early (22.39), on-time (24.71), and midway intervention (21.30), which did not differ significantly.

There was a significant effect of amount of increase, $F(1, 38) = 49.89$, $MSE = 825.37$, $p < .001$, $\eta_p^2 = .57$, with a higher mean for large (27.83) than for small (11.59). There was a significant main effect of rate of increase, $F(1, 38) = 30.97$, $MSE = 112.27$, $p < .001$, $\eta_p^2 = .45$, with a higher mean for steep rate (22.07) than for shallow rate (17.35).

These effects were qualified by two significant two-way interactions. There was a significant interaction between timing and rate of increase, $F(3, 114) = 5.62$, $MSE = 119.06$, $p < .01$, $\eta_p^2 = .13$. The main feature of the interaction is that there

Table 2. *Mean causal judgements and numbers of participants responding "yes" to Question 1, Experiments 1a, 2a, and 3a*

			Experiment		
Timing	*Amount*	*Rate*	*1a*	*2a*	*3a*
Early	Small	Shallow	14.3 (30)	21.0	20.8 (34)
		Steep	13.0 (28)	19.73	1.3 (35)
	Large	Shallow	31.7 (35)	40.34	8.6 (36)
		Steep	30.5 (34)	38.75	5.7 (39)
On time	Small	Shallow	11.8 (32)	28.3	19.9 (37)
		Steep	17.6 (36)	24.0	37.7 (39)
	Large	Shallow	31.8 (36)	44.2	64.0 (40)
		Steep	37.5 (38)	41.7	61.9 (39)
Midway	Small	Shallow	10.6 (28)	15.5	3.5 (11)
		Steep	16.7 (33)	15.9	3.2 (8)
	Large	Shallow	23.7 (30)	26.2	11.2 (12)
		Steep	34.1 (36)	35.6	10.2 (8)
Late	Small	Shallow	3.6 (13)	7.8	1.4 (4)
		Steep	4.9 (12)	9.6	1.9 (4)
	Large	Shallow	11.1 (17)	13.8	5.4 (4)
		Steep	22.1 (34)	15.7	4.1 (4)

Note: Numbers in parentheses are numbers of participants responding "yes" to Question 1 (Experiments 1a and 3a only). For Experiment 1a, $n = 39$. For Experiment 3a, $n = 40$.

was no significant effect of rate of increase with the early intervention, $F(1, 38) = 0.58$, $MSE = 108.75$. However, perhaps the most noteworthy feature is that the effect of rate of increase was significant at the late intervention, $F(1, 38) = 7.79$, $MSE = 191.95$, $p < .01$, $\eta_p^2 = .17$. Thus, even when the intervention occurred at a point where the increase phase had ended, causal judgements were still significantly higher if the increase had been steep than if it had been shallow. The causal question specifically asked whether the chemical caused an increase or not, so this cannot be interpreted as a judgement that the chemical halted an increase that had been in progress.

The other significant interaction was between amount of increase and rate of increase, $F(1, 38) = 4.66$, $MSE = 104.41$, $p < .05$, $\eta_p^2 = .11$. There were significant effects in accord with the respective main effects in all comparisons. The effect of rate of increase was greater when the amount of increase was large than when it was small.

Discussion

If judgements were guided by a principle of temporal contiguity, the highest causal judgements should have been found for the on-time intervention, which occurred at the start of the increase phase and was therefore temporally contiguous with the start of the increase. In fact, means for the on-time, early, and midway interventions were not significantly different. The result for the early intervention is not unduly surprising. People may be familiar with the idea that different drugs can take differing amounts of time to produce a noticeable effect (from seconds or minutes to weeks or months), so an effect delayed by a few hours may not be incompatible with their mechanism beliefs. The result for the midway intervention is rather more unexpected.

The midway intervention occurs midway through the increase phase: A noticeable increase has been underway for some hours and continues at the same rate for some more hours after the intervention. The intervention cannot have been responsible for initiating this increase, and there is no evidence that it changes the rate of increase. In short, there is no objective evidence that the midway intervention has any effect. Yet it was rated as high as the on-time intervention. It could be objected that the occurrence of an increase may not be clear halfway through the increase phase. If participants are detecting a signal amongst the noise of natural fluctuation in a dependent measure, perhaps the signal is not detectable until after the halfway mark in the increase phase. A glance at Table 1 is sufficient to refute that argument. In the baseline phase the count never varies by more than one for six consecutive time periods. In datasets with large amount of increase, the midway intervention occurs at a point where the count has risen by 5 from the first reading in the increase phase, and by 7 from the last reading in the baseline phase, in steady increments of 1 in the shallow rate datasets and of 2 in the steep rate datasets. It is surely not difficult to detect such an increase against a background of almost no variation. Nor is it difficult to detect that the rate of increase does not change after the midway

intervention, until the end of the increase phase. Participants identified an intervention as a cause of something that had unambiguously already started to happen before the intervention occurred.

It could be argued that the means were low—for the large amount of increase, between 30 and 40 on the 101-point scale—so most participants did not regard the chemical as a cause of the increase. However, the means were just as low for the on-time intervention. Whatever factors determine the generally low level of judgement, they apply equally to the on-time and midway interventions. More importantly, although the means were low, most participants did identify the chemical as a cause of the increase (see Table 2). For example, for the combination of midway intervention, large amount of increase, and steep rate of increase, only three participants (8%) responded "no" to the causal question. That is, more than 90% of participants identified the midway intervention as causal in that dataset.

There were, as predicted, significant effects of both amount and rate of increase. It was argued that both factors are indicators of the strength of causes, and the results are consistent with this reasoning. It is difficult to draw causal inferences with any confidence from data concerning a sample of one and lacking a control group. One way of dealing with these methodological imperfections is to rely on multiple cues to causality. If contingency alone is not a reliable guide, the conjunction of contingency, temporal contiguity, and outcome magnitude (both amount and rate of change), perhaps combined with the lack of explicit alternative causes (Einhorn & Hogarth, 1986) might make a subjectively more compelling case.

Why were the means low? The strong effect of amount of increase indicates that this could be the key factor. From a baseline mean of 15.5, the level rises only to 22 when the amount of increase is small and 27 when it is large. The respective means for these were 11.59 and 27.83. This indicates that a greater amount of increase, say to 47, would be likely to result in substantially higher judgement. It is also possible that a smaller baseline mean, say of 1.5, would result in higher judgement. An increase of 10 from a baseline of 1.5 is a 667% increase in the count, whereas an increase of 10 from a baseline of 15.5 is only a 65% increase. Percentage increase in the outcome could be a significant determinant of a judgement of causal strength. This will be tested in Experiment 4.

EXPERIMENT 1B

The design for Experiment 1a included both an unambiguous increase phase and a plateau in the postincrease phase, so that there was a substantial and consistent difference between baseline phase and postincrease phase readings. It might be thought that the occurrence of an unambiguous increase over a series of successive time periods is sufficient for causal inference, but in Experiment 1a this increase was confounded with the plateau state of the postincrease phase, and either could be important. In Experiment 1b the plateau was compared with a postincrease phase in which readings gradually declined to the baseline level. Specifically, four of the datasets from the design for Experiment 1a were compared with four datasets that were similar except for the occurrence of a decline in the postincrease phase.

Method

The method was similar to that in Experiment 1a, except for the design. The steep rate condition of Experiment 1a was used in all datasets here. Three variables were manipulated. Timing of intervention was manipulated with two of the conditions from Experiment 1a: early and on time. Amount of increase was manipulated with the same two values as those in Experiment 1a: small and large. The increase phase began in time period 7 and ended in time period 12. Following this, readings either remained at the same level, as in Experiment 1a, or declined to the baseline level. This decline was complete by time period 19, and readings remained at the baseline level (varying by one, as in the baseline period) until the end at time period 24. This condition is called the decline. The postincrease phase therefore had two conditions: plateau (four datasets from Experiment 1a) and decline. This was a

within-subjects manipulation. These manipulations yielded a total of eight datasets.

Result

Data were analysed with a 2 (timing of intervention: early vs. on time) × 2 (amount of increase: small vs. large) × 2 (postincrease phase: plateau vs. decline) within-subjects analysis of variance (ANOVA). Means are reported in Table 3.

The main finding of interest was a strong main effect of postincrease phase, $F(1, 38) = 38.45$, $MSE = 426.17$, $p < .001$, $\eta_p^2 = .50$, with a higher mean for the plateau (24.66) than for the decline (10.17).

There was a significant main effect of timing, $F(1, 38) = 5.82$, $MSE = 114.03$, $p < .05$, $\eta_p^2 = .13$, with a higher mean for the on-time intervention (18.87) than for the early intervention (15.96). This was qualified by a significant two-way interaction with postincrease phase, $F(1, 38) = 8.59$, $MSE = 75.90$, $p < .01$, $\eta_p^2 = .18$. This showed that the effect of timing was restricted to the plateau, $F(1, 38) = 12.61$, $MSE = 104.31$, $p < .001$, $\eta_p^2 = .25$, with a higher mean for the on-time intervention (27.56) than for the early intervention (21.76). There was no significant effect for the decline, $F(1, 38) = 0.00$, $MSE = 85.62$.

There was a significant main effect of amount of increase, $F(1, 38) = 44.58$, $MSE = 299.48$, $p < .001$, $\eta_p^2 = .54$, with a higher mean for large (23.96) than for small (10.87). This was qualified by a significant two-way interaction with postincrease phase, $F(1, 38) = 11.15$, $MSE = 219.09$, $p < .01$, $\eta_p^2 = .23$. There were significant effects in accord with the directions of the respective main effects in all analyses. It is likely that the interaction represents a floor effect, with the size of the differences noticeably reduced for means closer to the floor of the scale. This can be seen in Table 3. There were no other significant results.

Discussion

Replacing the plateau with a decline in the postincrease phase resulted in significantly lower causal judgements. In the decline datasets, Table 3 shows that in all cases more than half of the participants judged that the intervention was a cause of the change, even though the mean judged change was low. It may be that sustained improvement is necessary for an intervention to be judged a strong cause, though not necessary for it to be judged a weak cause. Experiment 4 was designed to investigate this further by manipulating the time of onset of the decline. Experiment 4 was also designed to test the possibility raised in the discussion of Experiment 1a that the low ratings could be attributed in part to the high baseline value. In Experiment 4, low and high baseline values were used.

Discussion of Experiments 1a and 1b

The first issue to consider is how much the results of Experiment 1 owe to the method used. The dependent measure asked participants, first, to say whether

Table 3. *Mean causal judgements and numbers of participants responding "yes" to Question 1, Experiments 1b, 2b, and 3b*

			Experiment		
Timing	*Amount*	*Postincrease*	*1b*	*2b*	*3b*
Early	Small	Plateau	13.0 (28)	19.7	31.3 (35)
		Decline	6.5 (20)	16.3	18.8 (26)
	Large	Plateau	30.5 (34)	38.7	55.7 (39)
		Decline	13.8 (26)	27.2	36.0 (29)
On time	Small	Plateau	17.6 (36)	23.9	37.7 (39)
		Decline	6.4 (23)	18.4	29.4 (30)
	Large	Plateau	37.5 (38)	41.7	61.9 (39)
		Decline	14.0 (26)	26.0	42.1 (32)

the intervention was a cause of increase in the level of cells in the patient's blood or not and, second, how strong a cause of increase they think the intervention was. It could be argued that this form of words is ambiguous. It could be taken as referring to the magnitude and rate of change that occurs after the intervention, or as a measure of confidence that the intervention did make a difference. If participants treated it as a measure of their confidence in their judgement that the intervention made a difference, that could explain the generally low mean ratings as showing some awareness of the uncertainties in the information. Experiment 2 was designed to test this by using a wording of the dependent measure that could not be interpreted as a confidence rating: a single question worded, "To what extent does the intervention cause an increase in the level of cells in the patient's blood?".

The stimulus information in Experiment 1 was presented in numerical form. It could be argued that people rarely have to make judgements about a series of numbers in everyday life. Experiences such as levels of pain are rarely registered numerically, although people may still have an appreciation of how the magnitude of the variable changes over time: They might say that they feel a bit more pain today than yesterday, for example. An alternative way of presenting magnitude information is in the form of a graph. Although people rarely have to make judgements about graphical information in everyday life, it could be argued that graphs present magnitude and change information more directly, and that this might therefore engage different processes perhaps closer to those of everyday life. Experiment 3 was designed to test this by presenting the same stimulus information as that in Experiment 1 but in the form of graphs.

EXPERIMENT 2

EXPERIMENT 2A

Method

The method was exactly the same as that in Experiment 1, except for the following differences. There were 40 participants, none of whom had participated in Experiment 1. The dependent measure used in Experiment 1 was replaced with a single question: "To what extent does [the chemical] cause an increase in the level of [the cells] in [the patient's] blood?".

Result

Data were analysed with a 4 (timing of intervention: early vs. on time vs. midway vs. late) × 2 (amount of increase: small vs. large) × 2 (rate of increase: shallow vs. steep) within-subjects ANOVA. The Tukey test was used for post hoc paired comparisons. Means for the 16 datasets are presented in Table 2. In brief, there were significant effects of timing of intervention and of amount of increase that resembled the corresponding results of Experiment 1a, but the effect of rate of increase was not significant.

There was a significant effect of timing of intervention, $F(3, 117) = 49.31$, $MSE = 318.96$, $p < .001$, $\eta_p^2 = .56$. Post hoc paired comparisons revealed a significantly lower mean for late intervention (11.74) than for early (29.96), on-time (34.57), and midway (23.29) intervention. The mean for the midway intervention was significantly lower than those for early and on time, which did not differ significantly.

There was a significant effect of amount of increase, $F(1, 39) = 74.00$, $MSE = 442.16$, $p < .001$, $\eta_p^2 = .65$, with a higher mean for large (32.04) than for small (17.74). The effect of rate of increase was not significant, $F(1, 39) = 0.27$, $MSE = 153.53$.

These effects were qualified by two significant interactions. There was a significant interaction between timing of intervention and amount of increase, $F(3, 117) = 7.17$, $MSE = 182.83$, $p < .001$, $\eta_p^2 = .16$. Simple effects analysis revealed significant effects in accordance with the respective main effects in all cases. The interaction shows that the effect of amount of increase tended to weaken as the timing of the intervention became later. It was, however, still significant even at late intervention ($p < .01$).

There was also a significant interaction between timing and rate of increase, $F(3, 117) = 3.11$, $MSE = 174.30$, $p < .05$, $\eta_p^2 = .07$. Simple effects analysis revealed significant effects of timing of intervention at both values of rate of increase, and these resembled the main effect. There was a significant effect of rate of increase at midway intervention, $F(1, 39) = 5.41$, $MSE = 180.37$, $p < .05$, $\eta_p^2 = .12$, with a higher mean at steep (25.76) than at shallow (20.82). This is the direction of difference found in the main effect of rate of increase in Experiment 1a. There were no other significant effects of rate of increase here.

EXPERIMENT 2B

Method

The method was exactly the same as that in Experiment 2a.

Result

Data were analysed with a 2 (timing of intervention: early vs. on time) × 2 (amount of increase: small vs. large) × 2 (postincrease phase: plateau vs. decline) within-subjects ANOVA. Means are reported in Table 3. In brief, there were significant effects of amount of increase and postincrease phase, which resembled those found in Experiment 1b. The effect of timing of intervention was not significant, though the difference between the means was in the same direction and similar in magnitude.

There was a significant effect of amount of increase, $F(1, 39) = 52.71$, $MSE = 291.38$, $p < .001$, $\eta_p^2 = .57$, with a higher mean for large increase (33.44) than for small (19.58). There was also a significant effect of postincrease phase, $F(1, 39) = 16.55$, $MSE = 397.46$, $p < .001$, $\eta_p^2 = .30$, with a higher mean for plateau (31.04) than for decline (21.97).

These two effects were qualified by a significant interaction, $F(1, 39) = 9.80$, $MSE = 169.40$, $p < .01$, $\eta_p^2 = .20$. There were significant effects in accord with the directions of the main effects in all analyses. The pattern resembles that found in the corresponding interaction in Experiment 1b. There were no other significant results.

Discussion

Most of the results of Experiment 2 replicated those of Experiment 1. In Experiment 2a there were effects of timing of intervention and amount of increase that resembled the corresponding results of Experiment 1a. In Experiment 2b there were significant effects of amount of increase and postincrease phase that resembled the corresponding results of Experiment 1b. The significant interaction between amount of increase and postincrease phase also resembled that found in Experiment 1b. It is noteworthy that the means tended to be low in this experiment, not obviously different from those found in Experiment 1. This suggests that the low means in Experiment 1 indicate genuine judgements of causal strength and not ratings of confidence that there is a causal relation. It is also noteworthy that most participants gave nonzero judgements to the late interventions, as in Experiment 1a. Thus, even though ratings of late interventions tended to be low, the late intervention was still regarded as causal.

Some differences in results should be noted. In Experiment 2a, the mean for the midway intervention was significantly lower than those for the early and on-time interventions, which was not the case in Experiment 1a. On the other hand, in Experiment 2b there was no significant difference between the early and on-time interactions, whereas there was such an effect in Experiment 1b. It seems plausible that there are weak effects of timing between the early, on-time, and midway interventions that reach statistical significance on some occasions and not on others. In Experiment 2a there was a significant effect of rate of increase only at the midway intervention, whereas the main effect of rate of increase was significant in Experiment 1a. Some details of interactions differed as well.

In conclusion, the hypothesis that judgements were affected by the wording of the dependent measure cannot be rejected because the results

were not precisely identical, but the similarities are such as to suggest that the manipulation of wording has not completely transformed the process of judgement.

EXPERIMENT 3

EXPERIMENT 3A

Method

The method of Experiment 3 was exactly the same as that in Experiment 1, with the following exceptions. There were 40 participants, none of whom had participated in either of the previous experiments. Instead of numerical presentation, graphical presentation was used. Time periods were indicated on the *x*-axis and level of cells on the *y*-axis. The readings were presented as a line graph. The timing of the intervention was indicated by a vertical line drawn through the graph in the appropriate place. A sentence underneath each graph reminded participants that the vertical line marked the time of the intervention.

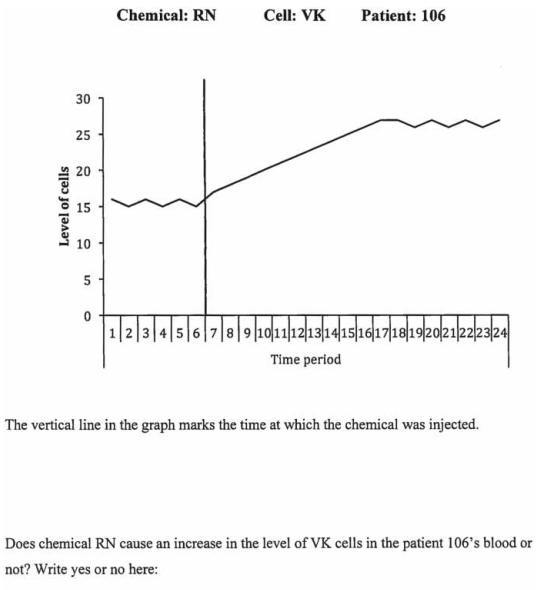

Figure 1. *Example stimulus presentation, Experiment 3.*

The instructions were modified accordingly. The passage referring to the columns of information (the latter part of the second paragraph of the instructions) was replaced with the following:

> You will see this information in the form of a graph. The time periods from 1 to 24 are indicated along the bottom and the level of cells up the side of the graph. The line across the graph from period 1 to period 24 shows how the level of the cells in question changed for that patient over the 24 time periods. A vertical line drawn on the graph from top to bottom shows you the time at which the chemical was injected.

Apart from that, the instructions were the same as those in Experiment 1. An example stimulus presentation is shown in Figure 1. The figure shows the on-time intervention, large amount, and shallow rate of increase presentation.

Result

Data were analysed with a 4 (timing of intervention: early vs. on time vs. midway vs. late) × 2 (amount of increase: small vs. large) × 2 (rate of increase: shallow vs. steep) within-subjects ANOVA. The Tukey test was used for post hoc paired comparisons. Means for the 16 datasets are presented in Table 2. In brief, there were significant effects of all variables that resembled those found in Experiment 1a, except that the means for the midway and late interventions were much lower than those for the early and on-time interventions. All four interactions were statistically significant. The main effects are reported first.

There was a significant effect of timing of intervention, $F(3, 117) = 137.37$, $MSE = 553.98$, $p < .001$, $\eta_p^2 = .78$. Post hoc paired comparisons showed that the means for the early (39.11) and on-time (45.88) interventions, which did not differ significantly, were significantly higher than those for the midway (7.06) and late (3.19) interventions, which did not differ significantly.

There was a significant effect of amount of increase, $F(1, 39) = 159.87$, $MSE = 312.87$, $p < .001$, $\eta_p^2 = .80$, with a higher mean for large (32.65) than for small (14.97) increase. There was a significant effect of rate of increase, $F(1, 39) =$

12.63, $MSE = 196.39$, $p < .001$, $\eta_p^2 = .24$, with a higher mean for steep (25.78) than for shallow (21.84) increase.

There was a significant interaction between timing of intervention and amount of increase, $F(3, 117) = 53.65$, $MSE = 163.71$, $p < .001$, $\eta_p^2 = .58$. This interaction reflects the greatly reduced judgements of the midway and late interventions. Thus, there were strong effects of amount of increase for early intervention, $F(1, 39) = 100.06$, $MSE = 273.10$, $p < .001$, $\eta_p^2 = .72$, and for on-time intervention, $F(1, 39) = 287.88$, $MSE = 161.45$, $p < .001$, $\eta_p^2 = .88$, both in accordance with the main effect. The effect of amount of increase at midway intervention was significant but smaller, and again in the direction of the main effect, $F(1, 39) = 10.52$, $MSE = 206.71$, $p < .01$, $\eta_p^2 = .21$, but the effect of amount of increase at late intervention was not significant, $F(1, 39) = 2.40$, $MSE = 162.74$.

There was a significant interaction between timing of intervention and rate of increase, $F(3, 117) = 5.45$, $MSE = 193.88$, $p < .01$, $\eta_p^2 = .12$. Here too the interaction reflects the low judgements of the midway and late interventions. There were strong effects of rate of increase for early intervention, $F(1, 39) = 16.36$, $MSE = 193.16$, $p < .001$, $\eta_p^2 = .30$, and for on-time intervention, $F(1, 39) = 8.55$, $MSE = 289.26$, $p < .01$, $\eta_p^2 = .18$, both in the direction of the main effect. However, there was no significant effect of rate of increase for either midway intervention, $F(1, 39) = 0.10$, $MSE = 160.18$, or late intervention, $F(1, 39) = 0.04$, $MSE = 135.43$.

There was a significant interaction between amount and rate of increase, $F(1, 39) = 8.97$, $MSE = 187.02$, $p < .01$, $\eta_p^2 = .19$. Simple effects analysis revealed significant effects of amount of increase at both rates of increase. There was an effect of rate of increase in accord with the main effect at small increase, $F(1, 39) = 26.56$, $MSE = 155.08$, $p < .001$, $\eta_p^2 = .41$, but not at large increase, $F(1, 39) = 0.17$, $MSE = 228.33$.

The three-way interaction was also significant, $F(3, 117) = 4.20$, $MSE = 194.49$, $p < .01$, $\eta_p^2 = .10$. This qualifies the interaction between amount and rate of increase. The effect of rate of increase is evident at early and on-time intervention, except for the combination of on-time intervention and large increase. This happens to be the combination that received the highest mean judgements (63.97 at shallow rate and 61.87 at steep rate), so possibly participants judged that rate of increase added nothing to what already appeared to be a strong effect of intervention.

EXPERIMENT 3B

Method

The method of Experiment 3b was exactly the same as that in Experiment 3a.

Result

Data were analysed with a 2 (timing of intervention: early vs. on time) × 2 (amount of increase: small vs. large) × 2 (postincrease phase: plateau vs. decline) within-subjects ANOVA. Means are reported in Table 3. In brief, there were significant effects of all three factors, which resembled those found in Experiment 1b.

There was a significant effect of timing of intervention, $F(1, 39) = 6.18$, $MSE = 691.41$, $p < .05$, $\eta_p^2 = .14$, with a higher mean at on-time (42.79) than at early intervention. There was a significant effect of amount of increase, $F(1, 39) = 155.05$, $MSE = 198.60$, $p < .001$, $\eta_p^2 = .54$, with a higher mean at large (48.94) than at small (29.32). There was a significant effect of postincrease phase, $F(1, 39) = 28.13$, $MSE = 647.98$, $p < .001$, $\eta_p^2 = .42$, with a higher mean at plateau (46.68) than at decline (31.59).

There was one other significant result, an interaction between amount of increase and postincrease phase, $F(1, 39) = 7.38$, $MSE = 233.66$, $p < .01$, $\eta_p^2 = .16$. This has the same pattern as the corresponding interaction in Experiment 1b.

Discussion of Experiments 1, 2, and 3

There is general consistency between the results of all three experiments. In all three, there were

significant and similar effects of timing of intervention, amount of increase, and postincrease phase. A main effect of rate of increase was found in Experiments 1 and 3 but not in Experiment 2, although a similar effect was found in Experiment 2 at midway intervention alone. An interaction between amount of increase and postincrease phase was found with a similar pattern in all three experiments.

There are indications, however, that the conditions peculiar to each experiment had some effect on judgements. I have already observed that the effects of timing of intervention varied to some degree between Experiments 1 and 2, although the variations could indicate the occurrence of a weak effect that does not always reach statistical significance. The results of Experiment 3 agree with those of the earlier two experiments in showing generally higher judgements for early and on-time intervention than for midway and late intervention. In Experiment 3, however, the means for midway and late intervention were substantially lower than those for early and on-time intervention, and a higher proportion of zero judgements occurred (see Tables 2 and 3). This suggests that the graphical format draws attention to the timing of the intervention in relation to the changes in levels of cells more than the numerical format does, and participants were more likely to realize that an increase that began before the intervention was unlikely to be the cause of it. It is not clear why this would be less likely to occur with the numerical format; this would be an interesting issue for future research.

One further possibility is that the results could be specific to the scenario, which was retained across all three experiments. A replication of the design of Experiment 1 has in fact been run using a different scenario. Participants were told that a gardener was trying to find out whether a fertilizer influenced the number of new leaves that plants produced in a given amount of time. The time scale was weeks instead of hours, but the numerical information was in other respects identical to that in Experiment 1. The results showed some differences in the details of significant interactions, but the main effects that are of central interest in this research were all replicated. Of particular note, there was no significant difference between the means for the early and midway interventions, though both were rated significantly lower than the on-time intervention. Replication with two different scenarios is not enough to conclude that there are no scenario-specific effects, but it is consistent with that possibility.

Mean judgements were generally higher in Experiment 3 than in Experiments 1 and 2. To illustrate, the highest mean judgement in Experiment 1 was 37.5 (on-time timing, large amount, steep rate), and the highest in Experiment 3 was 64.0 (on-time timing, large amount, shallow rate). Effect sizes were also generally larger in Experiment 3. To illustrate, the three effect sizes for the main effects in Experiment 1a were .28, .57, and .45; the corresponding effect sizes in Experiment 3 were .78, .80, and .24. One possible explanation is that participants in Experiments 1 and 2 had relatively low confidence in their judgements, perhaps because of the numerical format. In everyday life, time series information is unlikely to be presented in numerical format, but more likely to be an experience of intensity (of pain, for example), or some kind of visual information (e.g., from the gardening scenario just mentioned, the visible appearance of bushiness of a plant). Time series information is not likely to be available in graphical form either, but it is possible that graphical format is more intuitively related to intensity than numerical format is. In addition, the rise and fall of a line on a graph is very easily detected in visual perception, whereas the rise and fall in a set of numbers requires more deliberation and calculation to be assessed. The increases in means and effect sizes in Experiment 3 were not uniform, however. This is illustrated by the effect size for the main effect of rate of increase, which was .45 in Experiment 1a and .24 in Experiment 3a. This could be a further indication that judgements were generated by different processes, or it could indicate that the relative salience of different features of the information varies depending on presentation format. Further research could shed more light on this.

The after – before model

Experiments 1, 2, and 3 found effects of amount and rate of increase that would not be surprising: Higher causal judgements occurred for greater than for lesser amounts of change, and for faster than for slower change (although the latter was not found in Experiment 2). The effect of timing of intervention was less intuitive. In Experiment 1a, the on-time intervention was not rated significantly higher than the early or midway interventions. In Experiment 2a, the on-time intervention was rated significantly higher than the midway intervention, though not significantly higher than the early intervention. In both experiments, the midway intervention was consistently endorsed as a cause even though a steady incremental change had started before the intervention was made. Even more strikingly, a large proportion of participants made a nonzero judgement of the cause even with late intervention, an intervention made after the increase had stopped. For example, in Experiment 1a, with late intervention, large amount of increase and steep rate of increase, 87% of participants responded "yes" to the first question. The participants appear to have been judging that the intervention caused an increase that had actually finished by the time the intervention occurred. This was not found in Experiment 3a: Mean ratings of both midway and late intervention were much lower, and a majority of participants responded "no" to the first question.

In fact, a simple model can predict the results of Experiments 1 and 2 and most of the results of Experiment 3. Take the mean postintervention value and the mean preintervention value and subtract the latter from the former. I refer to this as the "after – before" model. Mean predicted values generated by this model for each of the main effects in Experiments 1, 2, and 3 are presented in Table 4, along with the means found in the experiments. Ordinal relations are of interest, not the absolute values. The table shows that the model does predict all the main effects in Experiments 1 and 2, except that it predicts a higher mean for midway intervention than for early intervention. The model also correctly predicts that the effect of amount of increase should be greater than the effect of rate of increase.

Most importantly, the model predicts the puzzling findings obtained for the midway and late interventions. The relatively high mean rating for the midway intervention is predicted by the model, as Table 4 shows. In addition, the model predicts an effect of amount of increase with late intervention. The predicted values are 4.19 for small and 7.95 for large. This difference is reflected in the observed means of 4.27 for small and 16.61 for large. The model also predicts an effect of rate of increase with late intervention, with means of 5.65 for shallow and 6.49 for steep. This difference

Table 4. *Predictions of the after – before model and main effects for Experiments 1, 2, and 3*

		Experiment			
I.V.	*Level*	*1*	*2*	*3*	*Predicted*
Timing	Early	22.39	29.96	39.11	6.25
	On time	24.71	34.57	45.88	6.94
	Midway	21.30	23.29	7.06	6.88
	Late	10.44	11.74	3.19	6.07
Amount of increase	Small	11.59	17.74	14.97	4.62
	Large	27.83	32.04	32.65	8.45
Rate of increase	Shallow	17.35	24.64	21.84	6.10
	Steep	22.07	25.15	25.78	6.97
Postincrease	Plateau	24.66	31.04	31.59	6.79
	Decline	10.17	21.97	46.68	2.90

was also found to be significant. This is particularly important because, on the surface, it appears as though judgements of the causal efficacy of the late intervention were affected by changes that occurred *before* the late intervention. If judgements were made in accordance with the after – before model, the effects really reflect the fact that the differences between the mean outcome values before and after the intervention are greater when amount of increase is large than when it is small, and when rate of increase is steep than when it is shallow. This explains what would otherwise appear to be a case of judged backwards causation, a cause affecting something that occurred earlier in time. It is possible that the finding is a genuine instance of judged backwards causation, and that a majority of participants did believe that the intervention could cause something that happened before it in time. However, the results of Experiment 3 indicate that this is unlikely: There, the means for the midway and late interventions were close to zero. If participants really believed in the possibility of backwards causation, it is unlikely that a mere change in presentation format would induce them to refrain from making judgements of that sort.

The main failing of the model is that it does not predict the low means for the midway and late interventions found in Experiment 3. This may show either that the model is wrong or that some other process is at work with the kind of graphical presentation that was used in Experiment 3. This is addressed in the general discussion. For now, the after – before model at least provides a suitable target, in that it is a parsimonious account that generates testable predictions. The remaining experiments were designed to run further tests of the model.

EXPERIMENT 4

Experiment 4 was designed to test two predictions of the after – before model. In the discussion of Experiment 1a, I commented that an increase from a low baseline mean was proportionately greater than the same absolute size of increase from a higher baseline mean. Participants might therefore make higher causal judgements for proportionately greater increases. The after – before model, however, predicts no effect of this manipulation because causal judgement depends only the absolute difference between the post- and preintervention means and should not, therefore, be affected by changes in proportions in the absence of changes in absolute differences. In addition, the model predicts an effect of timing of onset of the decline phase. The later the decline begins, the higher causal judgement should be, because the postintervention mean is higher when the decline starts late than when it starts soon after the intervention. To maintain methodological consistency with Experiment 1, these manipulations were set in a context of manipulations of timing of intervention and amount of increase in the increase phase.

Method

The method was similar to that of Experiment 1, with the following differences. There were 40 participants, none of whom had participated in the previous experiments. Four variables were manipulated. The basic design was similar to that of Experiment 1b with a decline occurring in the post-increase phase. The manipulated variables were the baseline outcome magnitude, the timing of the intervention, the amount of increase in the increase phase, and the timing of onset of the decline. Baseline outcome magnitude had two values, means of 5.5 (low) and 15.5 (high). In each case, baseline values differed by no more than one, as in Experiment 1. Timing of intervention had two values, early and on time, as in Experiment 1b. The early intervention came in time period 3 and the on-time intervention in time period 7. Amount of increase was measured as the arithmetical difference between the baseline mean and the value at the end of the increase phase. This took two values, 6.5 (12 – 5.5 in the low baseline condition and 22 – 15.5 in the high baseline condition; termed small), and 11.5 (17 – 5.5 in the low baseline condition and 27 – 15.5 in the high baseline condition; termed large). Timing of

Table 5. *Datasets used in Experiment 4*

Period	*l/s/13*	*l/s/17*	*l/s/21*	*l/l/13*	*l/l/17*	*l/l/21*	*h/s/13*	*h/s/17*	*h/s/21*	*h/l/13*	*h/l/17*	*h/l/21*
1	6	6	6	6	6	6	16	16	16	16	16	16
2	5	5	5	5	5	5	15	15	15	15	15	15
3	6	6	6	6	6	6	16	16	16	16	16	16
4	5	5	5	5	5	5	15	15	15	15	15	15
5	6	6	6	6	6	6	16	16	16	16	16	16
6	5	5	5	5	5	5	15	15	15	15	15	15
7	7	7	7	7	7	7	17	17	17	17	17	17
8	8	8	8	9	9	9	18	18	18	19	19	19
9	9	9	9	11	11	11	19	19	19	21	21	21
10	10	10	10	13	13	13	20	20	20	23	23	23
11	11	11	11	15	15	15	21	21	21	25	25	25
12	12	12	12	17	17	17	22	22	22	27	27	27
13	11	11	11	15	15	15	21	21	21	25	25	25
14	10	12	12	13	17	17	20	22	22	23	27	27
15	9	11	11	11	15	15	19	21	21	21	25	25
16	8	12	12	9	17	17	18	22	22	19	27	27
17	7	11	11	7	15	15	17	21	21	17	25	25
18	6	10	12	6	13	17	16	20	22	16	23	27
19	5	9	11	5	11	15	15	19	21	15	21	25
20	6	8	12	6	9	17	16	18	22	16	19	27
21	5	7	11	5	7	15	15	17	21	15	17	25
22	6	6	10	6	6	13	16	16	20	16	16	23
23	5	5	9	5	5	11	15	15	19	15	15	21
24	6	6	8	6	6	9	16	16	18	16	16	19

Note: Columns are identified by independent variables in the order baseline mean/magnitude/decline onset. For baseline mean, "l" = low and "h" = high. For magnitude, "s" = small and "l" = large. For decline onset, number indicates the time period in which the decline starts (13, 17, or 21).

onset of the decline was manipulated with three values, with decline starting in time period 13, 17, or 21. This yielded a total of 24 datasets. The design is shown in Table 5.

Results

Data were analysed with a 2 (baseline mean: small vs. large) × 2 (timing of intervention: early vs. on

Table 6. *Mean causal judgements and numbers of participants responding "yes" to Question 1, Experiment 4*

			Decline onset		
Baseline	*Timing*	*Magnitude*	*13*	*17*	*21*
Low	Early	Small	9.15 (21)	18.45 (32)	22.72 (32)
		Large	18.22 (29)	27.02 (34)	32.05 (36)
	On time	Small	15.05 (30)	20.45 (31)	24.52 (36)
		Large	25.62 (37)	34.20 (40)	35.57 (39)
High	Early	Small	8.82 (20)	15.72 (28)	18.80 (34)
		Large	24.72 (35)	32.67 (37)	32.95 (38)
	On time	Small	13.15 (27)	17.12 (33)	22.87 (34)
		Large	24.37 (34)	31.75 (38)	43.55 (40)

Note: Numbers in brackets are numbers of participants responding "yes" to Question 1 ($n = 40$).

time) × 2 (amount of increase: small vs. large) × 3 (decline onset: time period 13 vs. 17 vs. 21) within-subjects ANOVA. Means are reported in Table 6.

The results showed a clear effect of decline onset: The later the decline started, the higher the mean judgement was. For this effect, $F(2, 78) = 32.01$, $MSE = 350.93$, $p < .001$, $\eta_p^2 = .45$. Post hoc paired comparisons revealed the order 21 (29.13) > 17 (24.67) > 13 (17.39). There was no significant effect of baseline mean, $F(1, 39) = 0.04$, $MSE = 435.69$.

There was a significant effect of timing of intervention, $F(1, 39) = 14.16$, $MSE = 259.46$, $p < .001$, $\eta_p^2 = .27$, with a higher mean for on time (25.69) than for early (21.77).

There was a significant effect of amount of increase, $F(1, 39) = 80.39$, $MSE = 503.57$, $p < .001$, $\eta_p^2 = .67$, with a higher mean for large (30.22) than for small (17.24). This was qualified by a significant interaction with baseline mean, $F(1, 39) = 15.80$, $MSE = 102.37$, $p < .001$, $\eta_p^2 = .29$. Surprisingly for such a high F ratio, simple effects analysis shed little further light on this, because there was no significant effect of baseline mean at either large or small amount of increase. The range of means was greater for large than for small amount of increase, but this was not really reflected in either the respective F ratios or the effect sizes. For small, $F(1, 39) = 65.04$, $MSE = 199.24$, $p < .001$, $\eta_p^2 = .63$, and for large, $F(1, 39) = 71.65$, $MSE = 71.65$, $p < .001$, $\eta_p^2 = .65$. There were no other significant results.

Discussion

The results supported the predictions of the after – before model. There was, as predicted, no significant effect of baseline mean. This indicates that causal judgements are affected by absolute but not by proportional differences between pre- and postintervention means. There was a strong effect of timing of decline onset: As predicted, the later the decline started, the higher the mean causal judgement was. In addition, there were effects of timing of intervention and amount of increase that replicated those found in Experiment 1.

If people are computing unweighted means, this implies that they will be insensitive to the amount of time that passes after the intervention before a noticeable change occurs on the outcome measure. At the extreme, it would make no difference whether a given change occurred minutes or weeks after the intervention. This seems unlikely. While the exact effect of this temporal relation may depend on scenario-specific beliefs, or on cues to the duration of mechanisms connecting cause and outcome (Bühner & McGregor, 2006), it is likely that temporal contiguity matters, and that changes temporally proximal to the intervention will elicit higher causal judgements than changes more distal from it.

Experiment 5 was designed to test this by presenting a pattern of readings all at the same value but interrupted by a brief but readily noticeable change, which I call a "blip": a brief, sharp rise in the outcome value followed by an equally brief fall to the baseline value. This pattern of change could occur at any time after the intervention. If causal judgements are derived from unweighted means, that implies that the timing of the blip in relation to the intervention is irrelevant so long as the postintervention mean does not vary. On the other hand, if causal judgements are influenced by temporal contiguity, there should be higher causal judgements for blips that occur soon after the intervention than for blips that occur later.

EXPERIMENT 5

Method

Details of method were mostly the same as those for Experiment 1. There were 120 participants, none of whom had taken part in the previous experiments. There was one independent variable, timing of the blip, which was a between-subjects variable with three conditions, each with 40 participants. Apart from the blip, all observations, both pre- and postintervention, showed a value of 40. The blip comprised an increase by 8 per time period for three time periods, culminating in a maximum value of 64, immediately followed by a

decrease by 8 per time period back to the baseline value of 40. The intervention occurred after time period 8 in all three conditions. In the "early blip" condition, the blip began in time period 9, in the "middle blip" condition it began in time period 14, and in the "late blip" condition it began in time period 19. An example stimulus presentation is shown in the Appendix; that is the middle blip condition.

Results and discussion

Data were analysed with one-way ANOVA, and a significant result was found, $F(2, 117) = 5.01$, $MSE = 208.58$, $p < .01$, $\eta_p^2 = .08$. Post hoc paired comparisons with the Tukey test revealed that the mean for the early blip (18.70) was significantly higher than those for the middle (10.17) and late blips (9.55), which did not differ significantly.

The result therefore shows an effect of temporal contiguity, with higher causal judgements for the blip that was temporally contiguous with the intervention than for the blips that occurred after a delay. This indicates that a simple unweighted mean difference between pre- and postintervention values is not adequate as an account of causal judgements about time series.

GENERAL DISCUSSION

This research used a novel kind of presentation of stimulus information for causal judgement, taking the form of an individual case history with values of an outcome variable over a series of time periods combined with information about an intervention. Participants judged the extent to which the intervention caused an increase in value on the outcome variable. This format was designed to reflect some features that characterize information commonly available for causal judgement outside the laboratory, particularly the sample of one, the lack of any control or comparison group, and data about change over time. Some unusual results occurred: Principally, with numerical stimulus information, participants tended to judge that the intervention caused an increase even when an increase preceded the intervention and stopped at the time of the intervention. Causal judgements appeared to be influenced by several factors, including the amount of increase in the increase phase, the rate at which the increase occurred, and the timing of the intervention in relation to the increase phase. However, most of the results are predicted by a simple model in which causal judgement is generated by subtracting the preintervention mean from the postintervention mean. This suggests that participants are not operating with specific beliefs about what amount of increase, rate of increase, and the relation between change in the outcome value and the time at which the intervention occurs imply for causal judgement. Instead, they judge the efficacy of the intervention just from the difference between the values observed before and after the intervention, and the effects of the manipulated variables emerge from that.

There are, however, two results that are not predicted by the after – before model. Experiment 5 found evidence for an effect of temporal contiguity, in that higher judgements were found for a change that occurred immediately after the intervention than for changes that occurred later. As it stands, the after – before model makes no reference to temporal factors other than the division of the stimulus information into the periods before and after the intervention. This is likely to be an oversimplification, at best. The stimulus information used in this research has a limited span of 24 time periods, which are defined as hours in the scenario. The results of Experiment 5 indicate that these time periods are not equally weighted in computation of the after – before difference, and this is consistent with research showing that temporal contiguity is an important factor in causal judgement (e.g., Mendelson & Shultz, 1976; Michotte, 1963; Schlottmann, 1999; Shanks, Pearson, & Dickinson, 1989; but see also Bühner & McGregor, 2006). This implies that information about times more distal from the intervention, either before or after, would tend to carry progressively less weight in the computation, depending on the role of the judge's expectations about delay in occurrence of outcomes of interventions (Bühner

& McGregor, 2006). Eventually (and again depending on expectations) there would come a point where information carried zero weight because it was too far away from the intervention in time. The after – before model can be modified to account for the result of Experiment 5 by adding a weighting component, such that the postintervention mean is computed on observed values each multiplied by some weighting factor, with weights being maximal immediately after the intervention and declining thereafter. That would be a mere empirical fix, however, The finding shows that more than just integration of pre- and postintervention observations is involved in judgement, and it is likely that the finding reflects a general belief about temporal contiguity in causal processes. Further research on the way this is affected by scenario content factors, as in the study by Bühner and McGregor (2006) would be enlightening.

The second problematic result comes from Experiment 3, where the midway and late interventions received much lower judgements than the early and on-time interventions, which is not predicted by the model. Experiment 3 differed from the others in using graphical presentation instead of numerical presentation. It is likely that the result reflects a different approach to judgement with graphical information. One possibility is that different features of the stimulus information differ in salience between the two formats. It may be more obvious, with graphical presentation, that a steady increase has commenced before the midway intervention and that the rate of increase does not change after it. It is not clear why this should not be obvious with the numerical format, but if participants were treating preintervention and postintervention readings as separate blocks, then they may not have attended to the temporal dimension of change in relation to the timing of the intervention. Thus, even though the other results of Experiment 3 were predicted by the after – before model, it is likely that the low ratings of the midway intervention indicate that a different kind of process was involved, and that the after – before model does not fit. It is also possible that the same basic process operated, but that different kinds of information were weighted differently because of differences in attentive processing of the stimuli. Further research on the graphical presentation format should shed more light on this.

Normative causal judgement and practicality

One reason for investigating how causal judgements are made with contingency information about binary variables is that there are normative models of causal inference from such information, and the predictions of the normative models can be compared with observed judgemental tendencies (Bühner, Cheng, & Clifford, 2003). There can be no normative analysis of causality for the kind of information presented here, because it does not satisfy basic principles of experimental design (Cook & Campbell, 1979). There is just a sample of one and no control group.

There are normative analyses of interrupted time series data (Box, Jenkins, & Reinsel, 2008; Chatfield, 2004; McCain & McCleary, 1979), and there are several inferential techniques for deriving information such as effect size (e.g. Maggin et al., 2011; Parker, Hagan-Burke, & Vannest, 2007). Recently, Bayesian hypothesis testing methods have been developed for single subject designs (De Vries & Morey, 2013). The basic principle of Bayesian analysis is the rational modification of belief in the light of new evidence. Thus, the issue addressed by De Vries and Morey (2013) is the quantification of "the extent to which the data support the null hypothesis over the alternative hypothesis" (p. 168). This in turn can be regarded as indicating "the extent to which a rational person should adjust their beliefs, expressed in rational odds, in favour of the null hypothesis in response to the data" (p. 169). This, they argue, gives the Bayesian approach an advantage over traditional inferential statistics.

Be that as it may, all inferential tests rest on assumptions. De Vries and Morey (2013) noted that one of the assumptions of their method is that the observations that gave rise to the data were independent. This assumption is never met in time series data with a single subject because the process of observation itself can affect

subsequent readings, and of course successive readings may concern a process that unfolds continuously over the time series. No causal inference from interrupted time series data can ever be justified, not just because of the requirement of independent observations, but because of methodological shortcomings inherent in the interrupted time series design. The lack of a control group is an obvious example, and it means that the possibility of covert causal mechanisms that are correlated with the intervention cannot be ruled out. Cook and Campbell (1979) pointed out, for example, that causal inference from individual time series data is compromised by the possibility of cyclical processes such as those linked to time of day. A control condition would be the minimum requirement for dealing with this problem. The Bayesian approach could quantify the change in likelihood that the causal hypothesis is correct, but could never confirm a causal hypothesis, and the quantification would depend on how the methodological inadequacies of the design were taken into account.

Although it is not possible to specify what would be the right causal inference from time series data, it is possible to specify some causal inferences that would certainly be wrong. Judgement that the intervention causes an increase in the outcome variable when it occurs at the *end* of a change in the value of that variable, as was found in Experiment 1, is one of them. Whatever normative principles of causal inference might hold for data in the form of a time series, the causal judgements found in this research clearly do not conform to them.

Experiments 1 and 2 found that high proportions of participants made nonzero causal judgements about interventions that occurred after the increase phase had ended. It is difficult for any model in which human causal judgement is represented as normative to account for this because it would be incorrect, unless backwards causation really occurs. The only way for a normative account to deal with this result would be to show that participants reinterpreted the judgemental task in some way. For example, they might have opted to judge the extent to which the intervention was responsible for maintaining a high level on the outcome variable. This possibility cannot be ruled out on the basis of the present data, but it appears unlikely. First, a "maintenance" judgement is not compatible with the wording of the dependent measure, which explicitly asked participants to judge the extent to which the intervention caused an *increase*. Second, it does not explain why participants in Experiment 3 did not (with very few exceptions, shown in Table 2) give nonzero judgements for the late intervention: The wording of the judgement question was the same in Experiments 1 and 3, so it is not clear why participants would adopt a "maintenance" interpretation in Experiment 1 and not in Experiment 3. Third, it is not likely that people generally operate with a "maintenance" conceptualization of causality: the common assumption would be that things stay as they are unless acted on by a cause, which is contrary to the "maintenance" interpretation. Indeed, for exactly that reason it is not clear that a "maintenance" interpretation of causality could be defended on normative grounds. Hopefully further research investigating how people interpret the causal question will settle the matter.

Is human causal judgement adapted for methodologically ideal information or for practically obtainable information? I would argue for the latter. The conditions of life outside the laboratory rarely if ever satisfy basic methodological requirements. It may be possible to compare one group of observations with another, but it is not likely that the observations could be obtained under the controlled conditions required for valid causal inference. If I want to make a causal judgement about something to do with myself, such as recovery from a back injury, I cannot run a control condition in which I live a life that is exactly the same except for the absence of a hypothesized cause. Causal judgement must therefore be adapted (if at all) to conditions that prevail in the world because there is often a practical need to make causal judgements when information in normatively appropriate forms is not available. It is probably better to make causal inferences from methodologically inadequate information than to refrain from doing so, so long as the value or quantity of true

positives outweighs that of false positives. One way of improving the ratio of these would be to rely on multiple cues. Quantitative cues such as outcome magnitude and slope of incremental change could be combined with other relevant cues such as cues to mechanisms (Ahn, Kalish, Medin, & Gelman, 1995), generative transmission cues (Shultz, 1982), and contiguity.

Associative learning models

Models of associative learning have often been applied to human causal judgement (De Houwer & Beckers, 2002), and a test of one of them, the Rescorla–Wagner (R–W) model (Rescorla & Wagner, 1972) did much to stimulate research into human causal judgement from contingency information (Dickinson, Shanks, & Evenden, 1984). In application to human causal judgement, a possible cause is treated as a cue, and causal judgement reflects the strength of an associative bond between the cause and the outcome at the time of judgement. The bond tends to be strengthened by instances in which cue and outcome are both present and is weakened by instances in which the cue is present but the outcome does not occur.

The model holds for situations in which information is presented sequentially, in a series of learning trials. It might be thought, therefore, that the present format is well suited to the model. However, the problem is that all the stimulus information is available to the participant at the same time. Thus, the participant could generate a judgement by scanning back and forth through the information, computing means, adding, and subtracting, all facilitated by the presence of the entire dataset for judgement. This reasoning does suggest that the after – before model may be specific to the situation in which all the information is simultaneously available to the participant, and that it may be difficult or impossible to execute the computations required by the model when only one trial is available at a time. Investigating whether the after – before model predicts judgements when information is presented sequentially, and only one trial is before the participant's eyes at any one time, should therefore be a priority for future research, and in that situation the predictions generated by associative learning models could also be validly tested.

Issues and future directions

This first foray into causal judgements about information conforming to a simple interrupted time series design raises many issues for further work. Variance in the data was deliberately kept low in these experiments. The object was to ascertain the judgemental tendencies that would occur when the empirical data were unambiguous. That is, there should be clear demarcations between the baseline, increase, and postincrease phases. Participants were asked to judge whether the intervention caused an increase on the outcome variable, so it was important that the onset and offset of the increase be readily apparent in the materials. In addition, the increase phase was designed to show an incremental change, and the incremental nature of the change could be obscured if the variance in the data were increased. The mean value in the increase phase can be ascertained regardless of the amount of variance in the data, but it was thought that the occurrence of an incremental change might be more important than the mean during that phase. Variance is likely to be relevant, however, because, with a limited number of observations in each phase, detecting differences or changes is likely to be rendered more difficult by increasing variance in the data. Research on the applicability of signal detection theory to causal judgement with binary variables (Allan, Hannah, Crump, & Siegel, 2008) indicates that there may be much to learn from further investigation in this direction.

One of the methodological drawbacks with the interrupted time series design is the lack of a control group. It could be argued, however, that participants in these studies did have access to control group data because the manipulations were all within subject. Thus, they could compare one dataset with another. The instructions explicitly stated that each dataset concerned a different kind of cell, a different kind of chemical, and a different patient. Each was given a code identification at the

top of each dataset as a reminder that they were all different. It should have been clear to anyone cognisant of the need for a control group that the different datasets did not constitute control groups for each other. Using one dataset as a control for another would be like judging the effect of apples on liver function in one patient by comparing the data for that with data about the effect of oranges on heart rate in another patient: It would make no sense. If participants felt the need for control data, however, they might resort to other datasets in desperation. If that were the case, the results might show greater reluctance to endorse the "yes" response for the first dataset encountered than for later datasets, because for the first dataset no control data are yet available. I checked this for the data from Experiment 4. There were 28 "yes" responses to the first dataset, compared to a mean for all datasets of 25.75. The correlation between number of "yes" responses and ordinal position of dataset was −.24, $p = .3$. There is therefore no evidence that participants felt the need of control data before they were willing to endorse the causal hypothesis. Experiment 5 used a between-subjects design, so there was no opportunity for participants to use any other dataset as a control for the one being judged. Despite this, in the condition with the highest mean causal judgement, the early blip condition, 32 out of 40 participants (80%) gave a nonzero judgement. So the results give no indication that participants either used available datasets as control sets, or felt handicapped by the lack of alternative data.

I have made the point that there is no control condition with an interrupted time series design, but it could be argued that the baseline (preintervention) period is the control condition. However, comparison between preintervention as a control condition and postintervention as an experimental condition is only valid if nothing else changes in a way that is temporally correlated with the shift from preintervention to postintervention phase. The only way to ensure that this condition is met would be within the confines of a laboratory where all conditions can be held constant for the duration of the observations. Even then, in the case of a physiological scenario, it would be impossible to ascertain whether any physiological factors change in a way that is independent of but temporally correlated with the intervention. Cook and Campbell (1979) gave circadian rhythms as an example. Under the circumstances of everyday life, where control of variables is all but impossible, causal inferences from interrupted time series designs can never be justified.

The scenarios used in the present studies were designed to minimize the impact of preexisting causal beliefs by using unidentifiable cells and chemicals. It is nevertheless possible that judgements could have been influenced by causal knowledge in the form of expectations, either about the course of events in the increase phase given a certain pattern in the baseline phase, or about what would have happened in the postincrease phase if the intervention had not occurred. Causal judgements could then be influenced by assessment of the way in which the observed tendency deviated from the expected tendency. In the present research there is no information about participants' expectations. It is reasonable to suppose that participants would expect an apparently stable series of baseline measurements to continue in the absence of an intervention. What do they expect, however, if there is an increase phase prior to the intervention? This is more problematic. They are asked to judge the likelihood that the intervention causes an increase on the outcome variable. This appears to require that they judge, at the very least, that the postintervention observations are higher than they were expected to be in the absence of the intervention. If the increase is followed by a plateau, and the intervention occurs at the point where the increase levels off, endorsing "yes" must mean that the participant had an expectation that the increase would have reversed at that point if the intervention had not occurred. Since there is no sign of anything other than a steady increase in the periods leading up to the intervention, there is no reason for any participant to hold this expectation. Nevertheless it is possible to test the hypothesis by asking participants to extrapolate the readings they would expect to occur if the intervention had not been made. This might enable a discriminative test of this hypothesis and the after – before model.

A colleague objected that, because the participants were undergraduate students participating in return for course credit, they would adopt the easiest method of making the judgements. The problem with this is that there is no objective definition of easiness. Consider an analogy. If I am asked to solve $5 + 7$, then if I am experienced with mental arithmetic but have never used a pocket calculator before, mental arithmetic is easiest for me, and I would prefer to use that method to solve the problem. If I am used to using a pocket calculator but have never learned mental arithmetic, the calculator is the easier method for me, and I would preferentially use that. Thus, easiness can be defined in terms of whatever process is most natural for the person concerned. If the participants wanted to choose the easiest method and chose to compute the mean difference, that implies that that is indeed the easiest method for them, and it is easiest because it is the one with which they have had most practice. If they were more used to detecting trends and comparing the trends before and after the intervention, that would be the easiest method for them, and they would do that. So it can be argued that, even if the participants were opting for the easiest method, that is an indicator of which kind of judgemental process is most used by them.

The stimulus presentation format used in the present research was the simple interrupted time series design. There are many other quasiexperimental designs (Cook & Campbell, 1979), and it is likely that causally relevant information that is available to nonscientists often conforms to one or another of these. It is therefore important to investigate judgements made from stimulus presentations conforming to other kinds of quasiexperimental design and to investigate whether the after – before model is capable of predicting the results. Cook and Campbell (1979) argued that some quasiexperimental designs have a better claim to supporting causal inferences than others. It is likely that no causal inference from data conforming to a quasiexperimental design can be absolutely lacking in uncertainty; the arguments made by Cook and Campbell seem to focus mainly on reducing the likelihood of alternative explanations rather than eliminating them altogether. Nevertheless it would be important to investigate whether causal inferences are made with more confidence from some designs than from others, showing an intuitive appreciation of methodological issues relevant to causal inference. It is possible that the low causal ratings obtained in the present research reflect, in part, intuitive judgement of the limited value of the simple interrupted time series design for purposes of causal inference. If so, then higher ratings could be obtained from data conforming to more rigorous designs, such as designs in which appropriate control group data were available.

Outcomes don't just occur; they occur with a certain magnitude. Viewed from this perspective, the stimuli in causal judgement studies tend to be abstract or described: The participant is simply told that an outcome occurred or did not, and the magnitude information that would be available if a real outcome was encountered is missing. Further research on the way in which causal judgement is affected by outcome magnitude information is therefore important. The abstract nature of contingency information about binary variables typically presented in experiments makes it unrepresentative of causally relevant information typically encountered in the world, not just because of methodological considerations but also because of the absence of magnitude information. Yet the bulk of research on human causal judgement has used that kind of abstract, unrepresentative information. Investigating how causal judgement relates to outcome magnitude information, particularly in forms of information more likely to be available in the world, would give a different and arguably more faithful insight into how humans identify and judge the causes of outcomes that matter to them.

The present research is just a first step for this particular kind of causal judgement, but the after – before model can at least be a useful target for further attempts at disconfirmation. If these results do reflect real-world tendencies in causal judgement from time series information, then it seems likely that many causal beliefs are acquired

in non-normative ways from methodologically flawed information.

REFERENCES

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Allan, L. G. (1993). Human contingency judgements: Rule based or associative? *Psychological Bulletin*, *114*, 435–448.

Allan, L. G., Hannah, S. D., Crump, M. J. C., & Siegel, S. (2008). The psychophysics of contingency assessment. *Journal of Experimental Psychology: General*, *137*, 226–243.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). Hoboken, NJ: Wiley.

Bühner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119–1140.

Bühner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning*, *8*, 269–295.

Bühner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgment of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology*, *56A*, 865–890.

Bühner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology*, *57B*, 179–191.

Bühner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking and Reasoning*, *12*, 353–378.

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). London: Chapman and Hall.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

De Houwer, J. and Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *Quarterly Journal of Experimental Psychology*, *55B*, 289–310.

De Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*, 165–185.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology*, *36A*, 29–50.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*, 105–225.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.

Geddes, L. (2008, October 18). 'Roused from coma' by a magnetic field. *New Scientist*, 8–9.

Greville, W. J., & Buehner, M. J. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory and Cognition*, *35*, 444–453.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science: A Multidisciplinary Journal*, *31*, 765–814.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Jenkins, H. M., & Ward, W. C. (1965). Judgement of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*, 1–17.

Kotovsky, L., & Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. *Cognition*, *67*, 311–351.

Lagnado, D. A. & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 451–460.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic prior for causal learning. *Psychological Review*, *115*, 955–984.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalised least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, 49, 301–321.

McCain, L. J., & McCleary, R. (1979). The statistical analysis of the simple interrupted time-series quasi-experiment. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and Analysis Issues for Field Settings* (pp. 233–293). Boston: Houghton Mifflin.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, *26*, 209–239.

Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology*, *22*, 408–412.

Michotte, A. (1963). *The Perception of Causality*. New York: Basic Books.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education*, *40*, 194–204.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review*, *14*, 577–596.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*, 93–125.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, *35*, 303–317.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Tempral contiguity and the judgement of causality. *Quarterly Journal of Experimental Psychology*, *41B*, 139–159.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*, 1–51.

Shultz, T. R., & Kestenbaum, N. R. (1985). Causal reasoning in children. In G. Whitehurst (Ed.), *Annals of child development* (Vol. *2*, pp. 195–249). Greenwich, CT: JAI Press.

Siegler, R. S. (1975). Defining the locus of developmental differences in children's causal reasoning. *Journal of Experimental Child Psychology*, *20*, 512–525.

Siegler, R. S., & Liebert, R. M. (1974). Effects of contiguity, regularity, and age on children's causal inferences. *Developmental Psychology*, *10*, 574–579.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology*, *19*, 231–241.

White, P. A. (1988). Causal processing: origins and development. *Psychological Bulletin*, *104*, 36–52.

White, P. A. (2006). How well is causal structure from co-occurrence information? *European Journal of Cognitive Psychology*, *18*, 454–480.

APPENDIX

Example stimulus presentation, Experiment 5

Chemical: RF Cell: EJ Patient: 219	
Hour	*Level*
1	40
2	40
3	40
4	40
5	40
6	40
7	40
8	40
CHEMICAL INJECTED HERE	
9	40
10	40
11	40
12	40
13	40
14	48
15	56
16	64
17	56
18	48
19	40
20	40
21	40
22	40
23	40
24	40

Does chemical RF cause an increase in the level of EJ cells in the patient 219's blood or not? Write yes or no here:

If you said yes to the previous question, please rate how strong a cause of increase in the level of EJ cells in patient 219's blood chemical RF is: